A Note on the Weighted Means Problem

David V. Hinkley
University of Minnesota

Technical Report No. 315

*Research supported by National Science Foundation Grant MCS77-00959.

A Note on the Weighted Means Problem

David V. Hinkley

University of Minnesota

SUMMARY

Given several small normal samples with a common underlying mean, but without a common variance, how do we set confidence limits for the unknown mean when the different variances are unknown? A partially fiducial method of inference based on maximum likelihood is described which is amenable to simple normal approximation.

Key Words: Maximum likelihood; Fisher information; conditional inference; fiducial inference; ancillary statistics; curvature; non-homogeneous errors

1. INTRODUCTION

Suppose that we have several small sets of normal data with a common underlying mean μ but possibly different and unknown variances. The question then is how to weight the individual sample means to obtain a good overall estimate of μ; and, given such an estimate, how do we set confidence limits for the underlying mean? This is traditionally referred to as the "weighted means" problem, and has been of some interest to both theoreticians and practitioners of statistics. Theoretical interest arises because of the difficulty of eliminating the unknown variance parameters from inference about μ. Much work has focussed on the "simplifying" case where the number of data sets n is very large, while each set is itself small. Particularly important contributions are those of Bartlett (1936), and Neyman and Scott (1948), dealing with likelihood inference. Fiducial approaches to the problem have been discussed by Yates (1939) James (1956, 1959) and Fisher (1961a, b). Cochran (1954) and Cochran and Carroll (1953) discuss numerical approximations for setting confidence limits for μ using weighted means.

The present note suggests a "partially fiducial" interpretation of the model which removes the unknown variances. This leaves a non-homogeneous error model for the sample means, from which an exact conditional analysis is possible. We then extend results of Efron and Hinkley (1977) for large n to obtain a conditional normal approximation for the maximum likelihood estimator of μ. This leads to approximate conditional confidence limits for μ.

Section 2 contains some preliminary results concerning the likelihood function, and in Section 3 we review standard sampling properties of the maximum likelihood estimator of μ. Note is taken of the existence of approximately ancillary statistics. The partially fiducial model is discussed in Section 4.

2. Preliminary Results

The general formulation of the weighted means problem is as follows. The random variables Y_{jk} $(j = 1, \ldots, n,\ k = 1, \ldots, m_j)$ are independent such that Y_{jk} is $N(\mu, \tau_j)$, where $\mu, \tau_1, \ldots, \tau_n$ are unknown. The situation of interest is that where the m_j are small and n is fairly large. It is assumed that $m_j \geq 2$. We first reduce the data to the sufficient statistics

$$X_j = m_j^{-1} \Sigma_k Y_{jk} \quad \text{distributed} \quad N(\mu, \tau_j/m_j)$$

$$SS_j = \Sigma_k (Y_{jk} - X_j)^2 \quad \text{distributed as} \quad \tau_j \chi^2_{m_j - 1} \quad , \tag{1}$$

where χ^2_ν denotes a chi-squared variable with ν degrees of freedom. For convenience we shall often denote $(\tau_1, \ldots, \tau_n)$ by $\tau_{(n)}$, with corresponding meanings for $X_{(n)}$ and $SS_{(n)}$.

Based on model (1) the full log likelihood function $\ell_{\mu,\tau_{(n)}}(X_{(n)}, SS_{(n)})$ is

$$\ell = \text{const} - \tfrac{1}{2} \Sigma_j\, m_j \log \tau_j - \tfrac{1}{2} \Sigma_j\, m_j \tau_j^{-1} \{(X_j - \mu)^2 + SS_j\} \ . \tag{2}$$

Straightforward calculation shows that the Fisher information matrix for $(\mu, \tau_{(n)})$ is

$$\mathcal{I}_{(n)} = \text{diag}\, (\Sigma_j\, m_j \tau_j^{-1}, \tfrac{1}{2} m_1 \tau_1^{-2}, \ldots, \tfrac{1}{2} m_n \tau_n^{-2}) \quad , \tag{3}$$

whose inverse gives a lower bound for variances of unbiased estimators. Thus the lower bound for unbiased estimates of μ is $(\Sigma\, m_j \tau_j^{-1})^{-1}$, which is attained by the weighted mean

$$t(X_{(n)}, \tau_{(n)}) = \Sigma\, m_j \tau_j^{-1} X_j / \Sigma\, m_j \tau_j^{-1} \quad , \tag{4}$$

i.e. the maximum likelihood estimate for <u>known</u> $\tau_{(n)}$. Although the lower bound applies also when $\tau_{(n)}$ is unknown, $\mathcal{J}_{(n)}$ does not increase indefinitely as $n \to \infty$, so that standard asymptotic efficiency properties of maximum likelihood do not apply. (Thus asymptotic efficiency of the m.l.e. of one of several parameters does not follow from the fact that the corresponding element of the inverse information matrix tends to zero; neither does consistency follow from this!)

3. Maximum Likelihood: Sampling Properties

When the components of $\tau_{(n)}$ are unknown, the estimates

$$\bar{\tau}_j = SS_j/(m_j - 1) \qquad j = 1, \ldots, n$$

are not fully informative and their substitution in (4) gives an inefficient estimator. Indeed the latter estimator would be inconsistent as $n \to \infty$ if just one of the m_j were to equal 2. That $\bar{\tau}_j$ is not fully informative may be seen from (3), but on intuitive grounds one can see that information about τ_j is contained in $m_j(X_j - \bar{\mu})^2$ for $\bar{\mu}$ any estimator of μ. In fact if $\bar{\mu}$ is consistent, as is $\bar{X} = n^{-1} \Sigma_j X_j$ for example, then $\{SS_j + m_j(X_j - \bar{\mu})^2\}/m_j$ is asymptotically fully informative. The maximum likelihood estimator $\hat{\mu}$ takes cogniscance of this extra information in a fairly obvious way by substituting

$$\hat{\tau}_j^2 = \{SS_j + m_j(X_j - \hat{\mu})^2\}/m_j \qquad j = 1, \ldots, n$$

in (4), thus defining an implicit equation for $\hat{\mu}$, while suggesting also an iterative method for its calculation.

More formally, differentiation of the log likelihood (2) gives

$$\frac{\partial \ell}{\partial \mu} = \Sigma\, m_j \tau_j^{-1}(X_j - \mu)\ , \quad \frac{\partial \ell}{\partial \tau_j} = -\frac{m_j}{2\tau_j} + \frac{m_j(X_j - \mu)^2 + SS_j}{2\tau_j^2} ,$$

Then by substitution of the solution $\hat{\tau}_j(\mu)$ to $\partial\ell/\partial\tau_j = 0$ for fixed μ we have the equation

$$\left(\frac{\partial \ell}{\partial \mu}\right)_{\tau = \hat{\tau}(\mu)} = \Sigma \frac{m_j(X_j-\mu)}{SS_j + m_j(X_j-\mu)^2} = \frac{d\ell^*_\mu}{d\mu} , \qquad (5)$$

where $\ell^*_\mu = \ell_{\mu, \hat{\tau}(\mu)}$ is the maximum relative likelihood of Kalbfleisch and Sprott (1970). Another interpretation for ℓ^*_μ is given in Section 4.

Provided that all m_j exceed 2, there is a consistent solution $\hat{\mu}$ to $\frac{d\ell^*_\mu}{d\mu} = 0$ as $n \to \infty$, and the variance of the limiting normal distribution is

$$\Sigma \frac{m_j^2}{m_j - 2} \tau_j^{-1}/(\Sigma m_j \tau_j^{-1})^2 . \qquad (6)$$

This is dominated, however, by the estimator of Bartlett (1936) and Neyman and Scott (1948), which curiously ignores samples of size $m_j = 2$.

Previous analyses, such as those just mentioned, have been unconditional; thus (6) is an unconditional limiting variance. A conditional analysis requires the existence of an ancillary statistic, which here means a statistic distributed independently of $(\mu, \tau_{(n)})$. Thus the SS_j are not ancillary. However, if $\bar{\mu}$ is any consistent estimate, even $\bar{\mu} = n^{-1}\Sigma X_j$, then the statistics

$$Z_j = (X_j - \bar{\mu})/\sqrt{SS_j} \qquad j = 1, \ldots, n \qquad (7)$$

are asymptotically ancillary, and so provide the basis for an approximately efficient conditional analysis. Note that Z_j measures the discrepancy between two estimates of τ_j , namely those which are pooled in determining $\hat{\mu}$. This conditional analysis is algebraically complicated and has not been carried out here. It is clear that the resulting conditional distribution depends on the unknown variances $\tau_{(n)}$, as does the unconditional distribution.

The fact that the distributions of $\hat{\mu}$ and other estimators depend so heavily on $\tau_{(n)}$, which itself is poorly estimated when the m_j are small, is a serious limitation to the usefulness of such estimators, unless a different form of inference is used. There are, however, some useful practical approximations for use in interval estimation given by Cochran (1954).

4. Maximum Likelihood: Partially Fiducial Analysis

Fiducial analyses of the weighted means problem have been discussed by Yates (1939) and James (1956, 1959), among others. Somewhat related is the paper by Cox (1975) on partially Bayes inference. A direct way to obtain the "fiducial solution" is to recognise that given $\tau_{(n)}$, the fiducial distribution of μ obtained from $X_{(n)}$ is $N(\mu_f, \sigma_f^2)$ where μ_f is the maximum likelihood estimator (4) for known $\tau_{(n)}$, and $\sigma_f^2 = (\Sigma\, m_j \tau_j^{-1})^{-1}$. This conditional distribution is then averaged with respect to the joint fiducial distribution of $\tau_{(n)}$ based on $SS_{(n)}$. The exact calculation is clearly complicated, and will not be described further here. Instead we shall follow a "partially fiducial" argument which allows us to take advantage of approximate conditional theory for the maximum likelihood estimator $\hat{\mu}$.

The major difficulty with sampling theory inference via $\hat{\mu}$ is the potential variability of SS_j in hypothetical repetitions. In the author's view it seems more appropriate to recognize the fiducial interpretation about $\tau_{(n)}$ given the value of $SS_{(n)}$, namely that $\tau_j = SS_j/\chi^2_{m_j-1}$ $(j = 1, \ldots, n)$. This implies that the relevant model for $X_{(n)}$ is

$$X_j = \mu + C_j T_{m_j-1}, \qquad j = 1, \ldots, n \tag{8}$$

where $C_j = \{SS_j/m_j(m_j-1)\}^{\frac{1}{2}}$ and T_ν denotes a Student - t variable with ν degrees of freedom.

As a point of contact with the preceding discussion, notice that the log likelihood function based on (8) is the log likelihood ℓ_μ^* obtained from (2) by maximizing with respect to $\tau_{(n)}$.

The quantities in (7) are now exactly ancillary if $\bar{\mu}$ is <u>any</u> location statistic, that is one which transforms into $a\bar{\mu} + b$ when each Y_{jk} is transformed into $aY_{jk} + b$. It is this fact which leads from (8) to simple approximate inference procedures for μ based on $\hat{\mu}$, the latter being the maximum likelihood estimator for model (8); $\hat{\mu}$ is a solution to $d\ell_\mu^*/d\mu = 0$ as defined in (5). We first define $A_{(n)} = (A_1, \ldots, A_n)$ by

$$A_j = (X_j - \hat{\mu})/C_j \qquad j = 1, \ldots, n \tag{9}$$

with C_j as in (8). Then $(A_1, \ldots, A_{n-1}, \hat{\mu})$ is a one-one transformation of $X_{(n)}$, and $A_{(n)}$ is ancillary. Therefore we now argue conditionally on $A_{(n)}$, and a simple extension of Fisher's (1934) fundamental result gives the conditional density of $\hat{\mu}$ as

$$f(\hat{\mu}|A_{(n)}) = \exp\{\ell_\mu^*(X_{(n)})\}/\int \exp\{\ell_t^*(X_{(n)})\}dt \quad , \tag{10}$$

where of course $X_j = \hat{\mu} + C_jA_j$ $(j = 1, \ldots, n)$. The homogeneous case $C_1 = \ldots = C_n$, $m_1 = \ldots = m_n$ has been discussed at length by Efron and Hinkley (1977), who prove the following results:

<u>Lemma</u> As $n \to \infty$ the conditional density of $\hat{\mu}$ approaches the $N(\mu, 1/I^*)$ density where $I^* = (-d^2\ell_\mu^*/d\mu^2)_{\mu=\hat{\mu}}$ (11)

is the observed information. If $\mathcal{I}^*$ denotes the expected Fisher information

$$\mathcal{I}^* = E(-d^2\ell_\mu^*/d\mu^2) \ ,$$

then

$$\frac{\mathrm{Var}\{\hat{\mu}|A_{(n)}\} - 1/I^*}{\mathrm{Var}\{\hat{\mu}|A_{(n)}\} - 1/\mathcal{I}^*} = O_p(n^{-\frac{1}{2}}) \quad . \tag{12}$$

The coefficient of variation of I^* , that is the square root of $\mathrm{Var}\{I^*/\mathcal{I}^*\}$, is asymptotically the statistical curvature γ^* of ℓ^*_μ defined by Efron (1975). Finally, as $n \to \infty$

$$\mathrm{pr}\{2(\ell^*_{\hat\mu} - \ell^*_\mu) \geq d \mid A_{(n)}\} \approx \mathrm{pr}(\chi^2_1 \geq d) \quad . \tag{13}$$

It is reasonably clear that Efron and Hinkley's proofs of these results extend to the more general case of arbitrary C_j, m_j; we omit formal proof.

The practical interpretations of the Lemma are as follows: $1/I^*$ is a better approximation to the conditional variance of $\hat\mu$ than is $1/\mathcal{I}^*$, the difference being potentially greater when γ^* is high. For example, when $C_1 = \ldots C_n$ and $m_1 = \ldots = m_n = 2$, $\gamma^* = 1.6/\sqrt{n}$ which is large for $n \leq 25$, indicating that small values of (12) are quite likely; in general $\gamma^* = O(m^{-1}n^{-\frac{1}{2}})$ for $m_j \approx m$ large. The more important part of the Lemma is (13), which says that the standard large-sample chi-squared approximation for the log likelihood ratio is <u>conditionally</u> valid.

The accuracy of the first-order results above increases as I^* increases, reasonable numerical agreement being found for $I^* > 5$. The chi-squared approximation indicated by (13) is numerically better than that for $I^*(\hat\mu - \mu)^2$. Higher-order approximations, involving higher derivatives of ℓ^*_μ at $\mu = \hat\mu$, are easily obtained by expanding (10) in Taylor series beyond the quadratic term; see Efron and Hinkley (1977).

Straightforward evaluation of I^* from (5) gives

$$I^* = \Sigma\, m_j^2 \frac{SS_j - m_j(X_j - \hat\mu)^2}{\{SS_j + m_j(X_j - \hat\mu)^2\}} = \Sigma\, m_j^2 \frac{m_j - 1}{SS_j} \times \frac{m_j - 1 - A_j^2}{(m_j - 1 + A_j^2)^2}. \tag{14}$$

Note that I^* is ancillary under model (8), and may be viewed as containing most of the ancillary information for large n . The statistical curvature referred to in the Lemma is evaluated as

$$\gamma^{*2} = \frac{\Sigma \frac{m_j}{m_j+2} \{\frac{(m_j+1)(m_j^2+6m_j+12)}{(m_j-1)(m_j^2+10m_j+24)} - \frac{m_j}{m_j+2}\} c_j^{-4}}{(\Sigma \frac{m_j}{m_j+2} c_j^{-2})^2} . \qquad (15)$$

5. Comment

The approximate analysis described in Section 4 seems quite appropriate for very small samples provided there are a moderate number of them; the value of I^* indicates the suitability of the basic normal approximation. However, the data might contain evidence of structure among the τ_j , in which case the empirical Bayes approach of Cox (1975) would seem preferable. The unconditional sampling theory properties of $\hat{\mu}$ referred to in Section 3 account for properties of hypothetical data sets which seem to have little to do with a given set of "ordinary data." A study of the conditional analysis suggested in Section 3 might be illuminating.

Of course our discussion is likely to be irrelevant in the extreme case of two groups of size two.

This research was carried out while the author was visiting Imperial College under the auspices of a SRC Senior Fellowship, and was also supported by the National Science Foundation. Thanks are due to Professor D. R. Cox and his colleagues for their kind hospitality at Imperial College.

References

BARTLETT, M. S. (1936). The information available in small samples. Proc. Camb. Phil. Soc., 32, 560-6.

COCHRAN, W. G. (1954). The combination of estimates from different experiments. Biometrics, 10, 101-29.

COCHRAN, W. G. & CARROLL, S. P. (1953). A sampling investigation of the efficiency of weighting inversely as the estimated variance. Biometrics, 9, 447-59.

COX, D. R. (1975). A note on partially Bayes inference and the linear model. Biometrika, 62, 651-4.

EFRON, B. & HINKLEY, D. V. (1978). Observed versus expected Fisher information: assessing the accuracy of the maximum likelihood estimator. Biometrika, 65, (to appear in Part 3).

FISHER, R. A. (1934). Two new properties of mathematical likelihood. Proc. Roy. Soc., A, 144, 285-307.

FISHER, R. A. (1961a). Sampling the reference set. Sankhyā, A, 23, 3-8.

FISHER, R. A. (1961b). The weighted mean of two normal samples with unknown variance ratio. Sankhyā, A, 23, 103-14.

JAMES, G. S. (1956). On the accuracy of weighted means and ratios. Biometrika, 43, 304-21.

JAMES, G. S. (1959). The Behrens-Fisher distribution and weighted means. J. R. Statist. Soc. B, 21, 73-90.

KALBFLEISCH, J. D. & SPROTT, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion.) J. R. Statist. Soc. B, 32, 175-208.

NEYMAN, J. & SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. Econometrica, 16, 1-32.

YATES, E. (1939). An apparent inconsistency arising from tests of significance based on fiducial distributions of unknown parameters. Proc. Camb. Phil. Soc., 35, 579-91.